청소년을 위한 유해 웹영상 차단 시스템의 구현

이 은^{O}애 *, 정 명숙 *, 김 재건 *, 하 석운 *,
* 경상대학교 전자계산학과

Implementation of An Inappropriate Web_I mages Blocking System for Youth

Eun-Ae^{O}Lee *, Myung-Suk Jung *, Jae-Gun Kim*, Seok-Wun Ha*,
* Dept of Computer Science, Gyeong-Sang National University

요 약

인터넷이 활성화되면서 청소년에게 유해한 영상을 제공하는 사이트들이 급속히 범람하고 있으며, 이로 인해 청소년들의 정신 건강이 심각하게 훼손되고 있다. 본 논문에서는 청소년들이 접근하기 쉬운 유해 URL의 웹 문서에 대해 그 문서 내에 포함되어 있는 영상들의 유해성을 판별하여 유해 영상을 선택적으로 차단할 수 있는 시스템을 구현하여 제시한다. 유해 URL들에 대해 실험한 결과, 제안한 시스템의 효율은 full nudity의 경우에는 89.6%, 반라의 경우는 70.1%의 차단 효율을 나타내었으며, 얼굴 영상의 경우는 2%의 오판별이 있었다.

1. 서론

인터넷의 발달로 웹의 유용성은 날로 더해가고 있다. 그러나 한편으로는 청소년의 정서에 유해한 정보가 범람하고, 특히 누드 혹은 성적 행위에 관련된 영상이나 동영상의 급속한 침범을 통해서 청소년의 유해 사이트에 대한 접근 가능성이 증대하여 청소년의 정신 건강을 해치는 일이 많아지고 있다. CyberPatrol과 같은 청소년을 위한 보호 프로그램이 개발되고 있다.[1,2,3] 그러나 제공하는 이들의 교묘한 수법, 즉 유해 URL의 명칭이나 구성 세부 명칭으로는 유해 사이트임을 판별하기 어렵게 하는 수법들을 사용하는 데는 청소년을 보호하는 데 한계가 있다. 이러한 상황에서 영상의 이름을 유해하지 않은 것으로 가장한 사이트를 차단할 수 있는 유일한 방법은 그러한 유해 영상이 사용자 브라우저에 나타나기 전에 그 영상의 유해성을 판별하여 유해하다면 해당 영상을 보여 주지 않거나 해당 URL에 접근하는 것을 강제로 금지하게 하는 방법밖에 없을 것이다. 본 연구에서는 청소년이 호기심으로 입력하거나 그렇지 않다 하더라도 입력한 URL의 문서 내에 존재하는 영상이 유해 영상임을 자동으로 판별하여 해당 영상을 문서 내에 출력되지 않게 함으로써 청소년들의 유해 영상 접근을 차단할 수 있는 시스템을 구현하여 제시하고자 한다. 구현한 유해 웹영상 차단 시스템의 판별 시스템은 신경망으로 구성하였으며[4], 학습영상의 학습방법, 실제 영상의 유해 판단 방법, 시스템의 성능 검진 및 실험 결과 등을 나타내었다.

2. 유해 웹 영상 차단 시스템

2.1 유해 웹 영상 차단 시스템의 구성도

본 논문에서 구현하고자 하는 유해 웹 영상 차단 시스템의 구성도는 다음 그림 1과 같다. 이 시스템은 크게 사용자 웹 브라우저, 웹문서 분석 및 영상 추출, 영상 유해성 판별,유해 영상 차단의 네 부분으로 구성하였다. 사용자 웹 브라우저단에서는 사용자가 관심을

가지고 URL을 입력하는 창과 유해 여부를 판단하고 유해한 영상이 있다면 차단한 후에 해당 웹 페이지를 보여주는 웹 페이지 출력 창으로 구성되어 있다. 웹 문서 분석 및 영상 추출단에서는 웹 문서를 파싱하여 영상 파일들만을 추출하는 부분이다. 영상 유해성 판별은 파싱하여 추출한 영상들의 유해성을 판별하는 판별기와 판별 결과 해당 영상의 유해 여부를 결정하는 부분이다. 그리고 유해 영상 차단 부분은 웹 문서 내에 존재하는 영상들 중에서 유해한 것으로 판별된 영상은 문서에 포함시키지 않고 디스플레이 되도록 하는 부분이다.

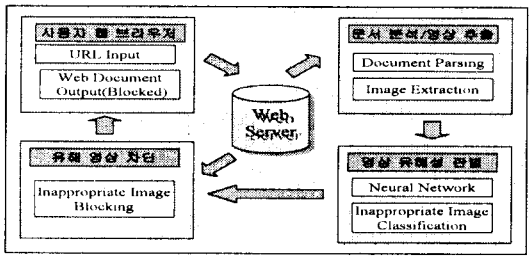

그림 1. 유해 웹영상 차단 시스템의 구성도

2.2 유해 웹 영상 차단 시스템의 구현

2.2.1 사용자 웹 브라우저

사용자가 유해 정보가 있는 URL을 입력하는 창과 유해 영상이 차단된 웹 문서를 출력하는 창으로 구성되어 있다. 다음 그림 2는 사용자가 관심을 두는 한 URL을 입력한 것을 보여주고 있다.

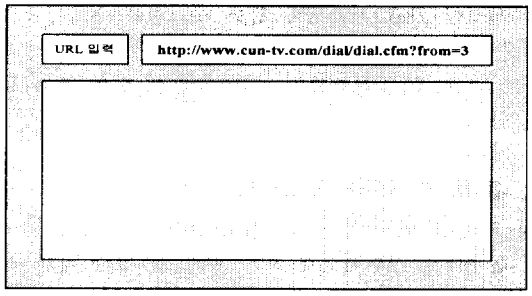

그림 2. URL 입력의 예

다음 그림 3은 입력 유해 영상이 차단되지 않았을 때의 URL의 웹 문서 예를 나타내었다.

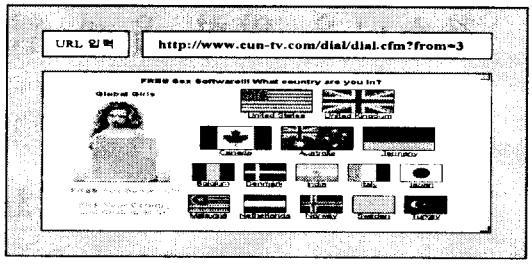

그림 3. 유해 영상이 차단되지 않은 경우의 URL의 웹 문서 예

다음 그림 4는 유해 영상이 차단된 웹 문서의 예를 나타내었다.

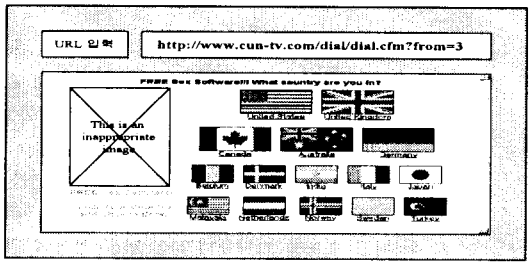

그림 4. 유해 영상이 차단된 웹 문서의 예

2.2.2 웹 문서의 분석 및 영상 정보 추출

웹문서를 분석하여 영상파일을 추출하는 과정은 JAVA 언어로 구현하였으며, 웹 문서 파싱을 행할 수 있는 객체 프로그램을 이용하였다.

2.2.3 영상 정보의 유해성 판별 신경망의 구현

2.2.3.1 영상 정보 유해성 판별 신경망의 구성

본 연구에서 구성하고자 하는 영상 정보 유해성 판별 신경망은 그림 5와 같이 다층 퍼셉트론 신경망으로 구성되었으며, 학습 알고리즘으로는 역전파 알고리즘을 사용하였다. 그리고 입력노드의 수는 입력 영상을 특징지우는 특징들의 종류 수와 같이 하였다. 특징으로는 입력 영상의 R, G, B 채널 각각의 평균 밝기, 각각의 밝기 분산 값, 붉은 정도와 푸른 정도 등이며, 이 특징들의 각각에 한 개의 뉴런을 대응시켜서 총 8로 구성하였다. 그리고 출력층의 노드 수는 유해하다 유해하지 않다에 대응하는 2개의 노드로 구성하였다.

또한 은닉층은 두 개의 층으로 구성하였으며, 은닉층 노드의 수는 400 개의 학습 패턴의 수에 대응시켜 제 1층에는 6개 제 2층에는 4개로 하였다. 은닉층과 출력층의 활성 함수는 모두 시그모이드(sigmoid) 함수를 사용하였다.

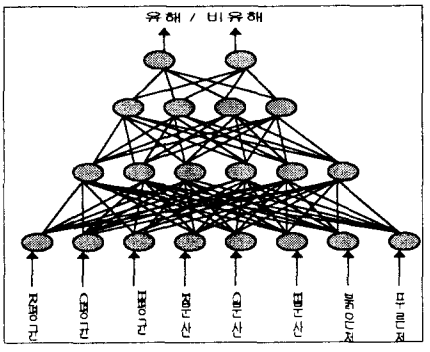

그림 5. 영상 정보 유해성 판별 신경망의 구성

2.2.3.2. 유해 및 비유해 학습 영상과 특징 추출

유해 학습 영상은 200개, 비유해 학습 영상은 200개를 사용하였으며, 유해 학습 영상의 예를 200개 중에서 100개만 그림 6에 나타내었으며, 그림 7에는 비유해 학습 영상 200개 중에서 100개만을 나타내었다.

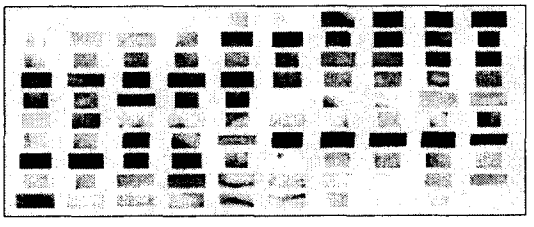

그림 6. 유해 학습 영상

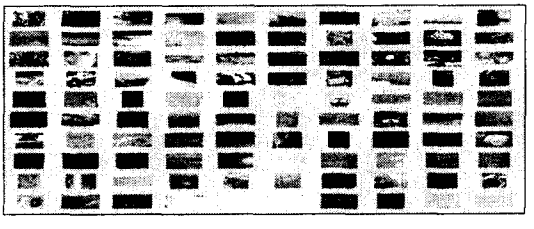

그림 7. 비유해 학습 영상

2.2.3.3. 학습 과정

각각의 학습 영상에 대해서 기대값(desired value)은 유해 학습 영상의 경우는 [1 0]으로, 비유해 학습 영상의 경우는 [0 1]로 하였다. 각각의 영상들에 대해서 R, G, B 채널의 밝기 평균을 구한다. *ver*는 학습 영상의 세로 길이, *hor*는 학습 영상의 가로 길이, $I(n,m)$은 채널별 픽셀의 밝기를 나타낸다.

$$R_{mean}=\frac{1}{ver*hor}\sum_{n=1}^{ver}\sum_{m=1}^{hor}I_R(n,m)$$
$$G_{mean}=\frac{1}{ver*hor}\sum_{n=1}^{ver}\sum_{m=1}^{hor}I_G(n,m)$$
$$B_{mean}=\frac{1}{ver*hor}\sum_{n=1}^{ver}\sum_{m=1}^{hor}I_B(n,m) \quad (1)$$

각각의 학습 영상들에 대해서 R, G, B 채널의 밝기 분산 값을 구한다.

$$R_{var}=\frac{1}{ver*hor-1}\sum_{n=1}^{ver}\sum_{m=1}^{hor}(I_R(n,m)-R_{mean})^2$$
$$G_{var}=\frac{1}{ver*hor-1}\sum_{n=1}^{ver}\sum_{m=1}^{hor}(I_G(n,m)-G_{mean})^2$$
$$B_{var}=\frac{1}{ver*hor-1}\sum_{n=1}^{ver}\sum_{m=1}^{hor}(I_B(n,m)-B_{mean})^2 \quad (2)$$

각각의 학습 영상들에 대해서 붉은 정도와 푸른 정도를 구한다. *RGB* 컬러 공간을 YC_bC_r 컬러 공간으로 변환한 다음, 각각의 영상에 대해서 그 푸른 정도와 붉은 정도의 평균 값을 구한다.

$$C_{b_{mean}}=\frac{1}{ver*hor}\sum_{n=1}^{ver}\sum_{m=1}^{hor}C_b(n,m)$$
$$C_{r_{mean}}=\frac{1}{ver*hor}\sum_{n=1}^{ver}\sum_{m=1}^{hor}C_r(n,m) \quad (3)$$

학습에 사용된 모든 영상들의 특징은 각 영상마다 $R_{mean}, G_{mean}, B_{mean}$과 $R_{var}, G_{var}, B_{var}$과 $C_{b_{mean}}$, $C_{r_{mean}}$의 여덟 가지이며, 이에 대응하는 신경망의 8개 입력 노드에 이들 값들을 차례대로 그리고 반복적으로 보여 주면서 기대값에 도달할 때까지 학습시킨다.

2.2.4 유해 영상 정보의 판별 및 차단

2.2.4.1 웹 영상의 분할

입력한 URL의 웹 문서 속에 있는 영상들의 유해성을

판별하기 위해서는 우선 웹 문서 속의 영상들을 일정한 크기로 분할하여야 한다. 여기서는 그 크기를 19*19으로 하였다. 그림 8, 그림 9, 그림 10은 각각 앞에서 제시한 URL의 웹 문서 석에 포함된 영상들 중에서 대표적인 세 가지 영상에 대해서 19*19의 크기로 분할한 웹 영상들을 각각 나타내었다.

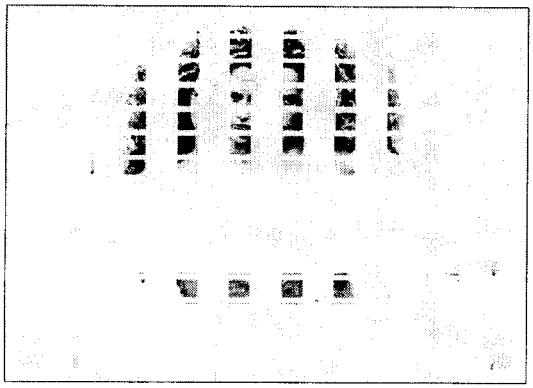

그림 8. 19*19의 크기로 분할된 웹 영상 1

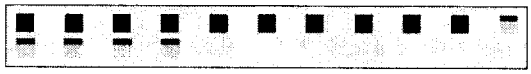

그림 9. 19*19의 크기로 분할된 웹 영상 2

그림 10. 19*19의 크기로 분할된 웹 영상 3

2.2.4.2 분할된 영상 조각들의 유해성 판별 과정

19*19의 조각으로 분할된 영상들의 8가지 특징들인 $R_{mean}, G_{mean}, B_{mean}$과 $R_{var}, G_{var}, B_{var}$과 $C_{b\,mean}$, $C_{r\,mean}$을 구한 다음, 이 특징 데이터를 영상 정보 유해성 판별 신경망의 입력으로 인가한다. 출력값의 한계는 다음으로 결정한다.

유해 영상의 경우에는 $[O_1>0.9\ O_2<0.1]$, 비유해 영상의 경우에는 $[O_1<0.3\ O_2>0.7]$의 범위에 속하는 때인 것으로 결정한다.

2.2.4.3 유해성 판별 결과

영상 1의 경우 총 181(17*11) 개의 입력 영상 중에서 유해 영상으로 판별된 것은 9개 비유해 영상으로 판별된 것은 나머지 172개이었다.

영상 2의 경우 총 15(3*5) 개의 입력 영상 중에서 유해 영상으로 판별된 것은 0개 비유해 영상으로 판별된 것은 나머지 15개이었다.

마찬가지로 영상 3의 경우에도 총 15(3*5) 개의 입력 영상 중에서 유해 영상으로 판별된 것은 0개 비유해 영상으로 판별된 것은 나머지 15개이었다.

따라서 영상 1의 경우는 유해 영상으로 판별되었으며, 영상 2와 영상 3은 비유해 영상인 것으로 판별되었다.

입력된 URL의 웹 문서에 포함되어 있는 모든 영상들에 대한 유해 여부를 판별한 결과를 다음 표 1에 나타내었다.

표 1. 입력된 URL의 웹 문서 내의 모든 영상들에 대한 유해 여부 판별 결과

영상 번호	분할영상 수	유해분할영상 수	비유해분할영상 수	유해여부
1	181	9	172	유해
2	15	0	15	비유해
3	15	0	15	비유해
4	6	0	6	비유해
5	6	0	6	비유해
6	6	0	6	비유해
7	6	0	6	비유해
8	6	0	6	비유해
9	6	0	6	비유해
10	6	0	6	비유해
11	6	0	6	비유해
12	6	0	6	비유해
13	6	0	6	비유해
14	6	0	6	비유해
15	6	0	6	비유해
16	6	0	6	비유해

3. 유해 웹 영상 차단 시스템의 평가

3.1 유해 웹 영상 차단 실험

유해 사이트에 대한 유해 영상의 판별 성능을 살펴보기 위하여 full nudity 경우의 100개 URL, half nudity 경우의 50개 URL, face image 경우의 50개 URL을 사용하였다. 이들 URL의 영상에 대한 유해 여부를 판별한 결과를 다음 표 2에 나타내었다.

표 2. URL의 영상에 대한 유해 여부 판별 결과

영상형태	URL 수	유해 영상 수	유해 영상에 대한 유해 판별 수 및 비율	유해 영상에 대한 비유해 판별 수 및 비율	판별 오류 수 및 비율	비유해 영상 수	비유해 영상에 대한 비유해 판별 수 및 비율	비유해 영상에 대한 유해 판별 수 및 비율	판별 오류 수 및 비율
full nudity (전라)	100	374	335 (89.6%)	13 (3.5%)	26 (6.9%)	112	108 (96.4%)	0 (0%)	4 (5.6%)
half nudity (반라)	50	87	61 (70.1%)	16 (18.4%)	10 (11.5%)	35	32 (91.4%)	0 (0%)	3 (8.6%)
face image (얼굴)	50	0				84	76 (90.5%)	2 (2.4%)	6 (7.1%)

3.2 실험 결과 및 평가

유해 사이트에 대한 유해 영상 차단 실험 결과, 유해 영상 조각들의 단순 학습에 의해서도 차단 효과가 뛰어남을 알 수 있었다. 전라(full nudity)인 경우와 반라(half nudity)인 경우는 그 차단 성능이 우수하였다. 반면에 수영복 차림이나 핫 팬티 차림 등의 경우에도 차단되는 예가 발생되었다. 그러나 사용자가 입력하고자 하는 URL이 이미 유해 사이트일 가능성이 다분하다는 점에서 차단의 당위성을 인정할 수 있을 것이다. 얼굴만이 노출되어 있는 영상에 대해서도 차단될 가능성이 있을 것으로 의심이 가지만 실제 얼굴의 경우 코나 입술 눈 등의 부위의 돌출성에 의해서 뚜렷한 명암이 존재하는 점을 감안할 때 실제 웬만큼 큰 얼굴 영상을 제외하고는 크게 문제되지 않는 것을 실험을 통해서 알 수 있었다.

4. 결론 및 향후과제

유해 웹영상 차단 시스템을 신경망을 이용하여 구현하였으며, 유해 영상 차단 실험을 통해서 바람직한 결과를 얻을 수 있었다. 반라의 경우에 대한 상세 분석과 얼굴 영상에 대한 상세 분석, 그리고 웹 상에서의 실시간 처리를 고려하여야 하겠으며, 향후 익스플로러나 넷스케이프와 같은 웹 브라우저 상에서의 구현과 일반적인 URL에 대한 유해 여부 실험을 행할 것이다.

[참고문헌]

[1].Access Management Engine(AME).
http://www.bascom.com
[2].AltaVista Filtered Search Service.
http://www.altavista.digital.com
[3].Cyber Patrol.
http://www.learningco.com
[4].김회승. "영상인식-영상처리,컴퓨터 비젼,패턴인식, 신경망-".생능출판사